Article

Int J Organ Transplant Med 2026; 17(2), 17–32

From Risk Prediction to Clinical Review Queues: Multicenter Alert Stewardship for Mortality and Graft-Loss Risk After Kidney Transplantation

Seyed Asad Alireza[1,*], Seyed Alireza Taghavi[1]

[1] *Gastroenterohepatology Research Center, Shiraz University of Medical Sciences, Shiraz, Iran*

[*] *Correspondence: asadalir@sums.ac.ir*

ABSTRACT

Background: Aggregated dynamic models for kidney transplant follow-up learning are generally described by AUROC, AUPRC, threshold precision, and feature importance. These characteristics are needed for model assessment, but they fail to answer the question of an operationally relevant deployment decision faced by the transplant network: how many patients will enter the review process per center, how many events will be captured, and what additional workload will be incurred through greater recall? **Objective:** The aim of this study was to find out if aggregate prognosis tables for annual death and graft-loss risk can be converted into a center-specific review ledger, which will reveal an operationally sound choice of an alerting threshold without creating a new predictive model and a new optimization procedure. **Methods:** Decision-Calibrated Domain Balancing (DCDB) and Multicenter Alert Stewardship Augmentation (MASA) were applied to the numeric tables generated by the STCS to convert them into review queue metrics. DCDB calculated prevalence-adjusted enrichment, longitudinal gain, workload, and domain saturation. MASA recovered the follow-up prevalence from recall, precision, and specificity; estimated flagged patients, captured events, and false positives; disaggregated alert counts by six center denominators; quantified threshold expansion; and summarized the effective predictor-domain breadth. **Results:** The deployment ledger included a longitudinal panel of model–outcome pairs (20 rows); an operating point panel (10 rows); a 60-row center–outcome–recall panel of alerts; and a domain panel (9 rows). LightGBM demonstrated the best graft-loss enrichment ($\rho = 18.52$); the death prediction model showed lower enrichment and faster workload growth. Follow-up prevalence recovery showed no threshold dependency, with the averages 2.82% for death and 1.51% for graft loss. A death recall rule of 0.50 will flag approximately 651 patients and capture around 68 events. A graft-loss recall rule of 0.60 will flag approximately 271 patients and capture 43 events; raising the graft-loss recall to 0.70 will bring the queue up to 516 patients capturing about 50 events. Zurich and Basel will generate the biggest review queues due to their large patient cohort denominator. Effective domain breadth was 2.51 for the pre-transplant inventory and 2.90 for the follow-up inventory, revealing that the nominal domain numbers overstate the explanatory power of the predictor domains. **Conclusion:** Aggregate prognosis tables can be converted into an actionable deployment case when the statistical performance is evaluated with regard to workload, center scale, and domain concentration. The study answered the question posed by providing a choice of a network-wide starting rule – graft-loss recall 0.60. Graft-loss recall 0.70 is a capacity-driven escalation, while broad death alerting calls for a stricter clinical filter prior to the deployment.

KEYWORDS: kidney transplantation; graft loss; death prediction; alert stewardship; precision-recall analysis; clinical workload; model deployment; domain concentration; machine learning

1 Introduction

Kidney transplantation is distinguished from dialysis through a period of structured surveillance rather than through a definitive endpoint of care. Post-surgical risk evolves depending on the stabilization or deterioration of renal function, the appearance of proteinuria, the occurrence of rejection episodes, the presence of donor-specific antibodies, the acquisition of infection and malignancy risk, and the manifestation of cardiovascular or metabolic disease. Chronic allograft loss is thus not a one-path process. Antibody-mediated damage is one pathway of graft injury [1]. Recurrent disease, toxicity, nonadherence, vascular disease, donor-related factors, infection, and chronic kidney disease are additional contributors to chronic allograft loss [2, 3]. Patterns of long-term allograft loss have been described across European kidney transplant populations [4], and prediction systems for allograft loss risk have been validated [5]. Similarly, mortality after transplantation varies over time due to changes in recipient age, renal function, cardiopulmonary disease, infection history, treatment history, and functional capacity. A clinically meaningful prediction system would thus need to allow for reevaluation over time as opposed to risk prediction only at transplantation.

More recent transplant studies have used machine learning to analyze allograft survival, patient survival, and longitudinal allograft function. Survival-statistical and Bayesian decision-support models have been tested in multicenter kidney transplant cohorts [6, 7]. Machine learning approaches for allograft survival prediction and feature significance analysis have been developed [8], while sequence-to-sequence deep learning has been applied to forecasting individual allograft function [9]. Recent efforts have included prospective modeling procedures and methods for interpretable prediction of graft and patient outcomes [10, 11]. This literature demonstrates that structured clinical data contains useful prognostic signal, but it also highlights the dependence of model performance on the time period under investigation, the availability of features, the case-mix, the endpoint definition, and center practices. Clinical decision support has accordingly focused on transparent incorporation into transplant workflows [12]. Reproducibility across transplant eras is an ongoing concern [13], and explainable machine learning has been applied to increase the interpretability of graft-survival predictions [14]. In the context of kidney transplantation, this concern is especially significant, since a risk score serves not merely to categorize a record, but also to instigate nephrologist review, transplant coordinator follow-up, lab workup, immunologic testing, medication assessment, or multi-disciplinary discussion.

The Swiss Transplant Cohort Study (STCS) offers a particularly pertinent environment for analyzing this deployment question. The STCS is a nationwide longitudinal cohort of solid organ transplant recipients in Switzerland, with kidney transplant data including donor and recipient characteristics, immunologic information, laboratory results, rejection episodes, infections, comorbidities, center indicators, and follow-up outcomes [15]. These data have been analyzed with regard to pre-transplant donor-specific antibodies [16] and early post-transplant systemic inflammation [17]. Fan et al. have used the STCS for developing a dynamic two-stage approach for kidney transplant recipients with a pre-transplant model for first year outcomes and an annual follow-up model for next-year death and allograft loss [18]. As seen in the numerical tables, follow-up data enhanced predictive performance, LightGBM delivered the highest graft loss performance, and TabPFN and LightGBM performed similarly with regard to mortality prediction. The same tables provide threshold operating characteristics at recall targets ranging from 0.50 to 0.90, providing a direct foundation for an implementation-oriented question beyond model comparison.

Standard reporting quantities are not by themselves sufficient to answer this implementation question. AUROC measures pairwise discrimination at all possible threshold points, while AUPRC is more sensitive to positive class ranking in imbalanced data [19, 20]. Calibration remains fundamental to prediction-model development [21]. Decision-curve analysis provides a framework for assessing clinical utility [22], while sample size considerations ensure reliable modeling [23]. Reporting standards like TRIPOD and PROBAST focus on transparent description of cohort variables, outcomes, modeling procedure, performance, and applicability [24, 25]. However, a transplant program needs to interpret a threshold as a clinical practice rule. A recall target decides the number of patients to be followed up, the number of actual events to be detected, the number of false reviews, and the impact of the same statistical

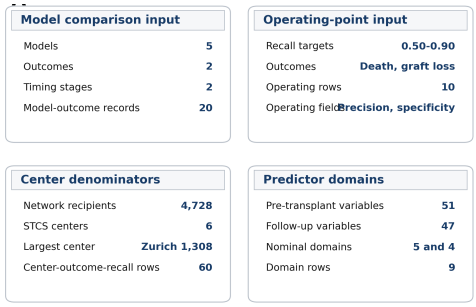
threshold in terms of differing absolute workload at different centers.

The aim of this paper is to answer the following research question: can threshold tables from a dynamic kidney-transplant prediction study be translated into a center-level review ledger to identify a feasible starting rule for next-year death and graft-loss surveillance? This paper does not intend to develop a new prediction model, new feature engineering technique, or new threshold optimization method. It uses aggregate threshold tables from the STCS combined with Decision-Calibrated Domain Balancing (DCDB) and Multicenter Alert Stewardship Augmentation (MASA) as post hoc translation layers. DCDB evaluates prevalence-adjusted enrichment, longitudinal gain, evaluation burden, and predictor domain saturation. MASA estimates event prevalence based on recall, precision, and specificity; computes alert volume and detected events; distributes alerts per center denominator; assesses threshold expansion; and computes effective predictor domain breadth.

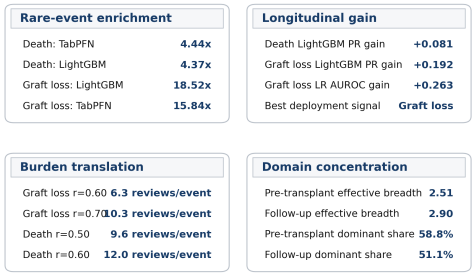
This paper thereby makes a specific and operational contribution. It shows how numbers from the STCS summary tables can be used to determine whether a statistically attractive threshold is clinically manageable in a multicenter transplant network. This is not an attempt at prediction-model development aimed at higher AUROC or algorithm innovation. This is rather an attempt at deployment interpretation: which outcome, prediction model signal, recall target, and center-level review workload define the most conservative starting rule for a real-world transplant follow-up service.

The structure of the analysis is visually summarized in Figure 1. The four panels separately present numerical inputs, DCDB quantities, MASA workload quantities, and finally candidate rules. This structure ensures that the analysis stays connected to the actual numbers from summary tables, while emphasizing that the main object of this paper is the review queue created by a threshold, not the theoretical model-performance ranking.

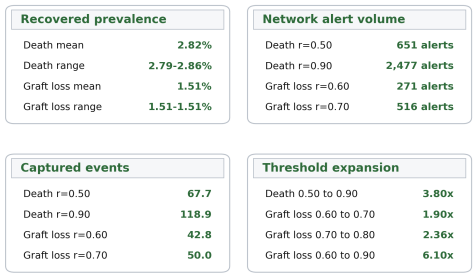
The four-panel output constitutes the audit trail. In Panel 1, we learn which aggregate tables form the evidential basis of our study. In Panel 2, the normalization of the statistical signal is explained for rare events. Panel 3 explains how thresholds get translated into expected workload, and Panel 4 provides the conversion of the quantities in Panel 3 into a set of clinical rules. This process ensures transparency because all conclusions have a specific STCS number behind them.



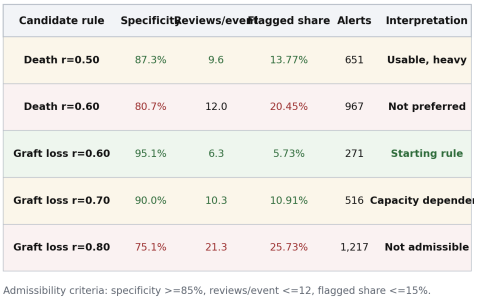
(a) Input tables.



(b) DCDB quantities.



(c) MASA quantities.



Candidate rule	Specificity	Reviews/event	Flagged share	Alerts	Interpretation
Death r=0.50	87.3%	9.6	13.77%	651	Usable, heavy
Death r=0.60	80.7%	12.0	20.45%	967	Not preferred
Graft loss r=0.60	95.1%	6.3	5.73%	271	Starting rule
Graft loss r=0.70	90.0%	10.3	10.91%	516	Capacity dependen
Graft loss r=0.80	75.1%	21.3	25.73%	1,217	Not admissible

Admissibility criteria: specificity >=85%, reviews/event <=12, flagged share <=15%.

(d) Deployment rules.

Figure 1. Analytical anatomy of multicenter alert stewardship.

2 Materials and Data Construction

2.1 Numerical material

The numerical material included the aggregate STCS tables for multicenter kidney-transplant machine-learning prediction [18]. Patient-level data was neither collected nor reconstructed nor re-identified. Fan et al. analyzed next-year mortality and graft-loss predictions using two data periods: pre-transplant for first-year risks based on donor-recipient information and annual follow-up for next post-transplant-year risks based on pre-transplant and yearly post-transplant observations. The latter corresponds to the routine annual follow-ups of transplantation and does not require equal-length sequences for each patient.

Analytic materials included four table families. The first table presented AUROC and AUPRC values for five algorithms: logistic regression, support vector machine, multilayer perceptron, LightGBM, and TabPFN. They were specified for the death and graft loss in pre-transplant and annual follow-up periods. The same table reported incidence proportion values of 2.61% for pre-transplant death, 2.79% for follow-up death, 4.07% for pre-transplant graft loss, and 1.49% for follow-up graft loss. The second table showed follow-up operating points for death and graft loss at recall targets of 0.50, 0.60, 0.70, 0.80, and 0.90 with respect to threshold, precision, specificity, balanced accuracy, and F1 score. The third table listed patient and donor characteristics for 4728 kidney transplant recipients in Bern, Lausanne, Geneva, St. Gallen, Basel, and Zurich. The fourth table reported number of features in the pre-transplant and annual follow-up domains of predictors.

Longitudinal tables in the STCS show that the at-risk population was shrinking with post-transplantation time, from more than 4000 in year one to fewer than 100 in year 13 with adverse events accumulation and uncertainty of the later years due to smaller denominators [18]. Thus, the alert ledger interprets the network-wide alerts as a planning denominator for the cohort, not as an assertion that all 4728 patients were eligible for alerting at a particular point in time. It is helpful in the context of deployment planning due to the presence of center denominator in the center table.

2.2 Analytical tables

The analysis dataset was structured to have four related panels. The longitudinal model panel had 20 rows, which accounted for five models, two outcome measures, and two timing periods. The longitudinal model panel recorded AUROC, AUPRC, longitudinal gain, and prevalence-adjusted enrichment. The operating-point panel had 10 rows, which accounted for two outcome measures and five recall rates. This panel recorded threshold, precision, specificity, balanced accuracy, and F1-score, and also recorded recovered prevalence, marked rate, network alerts, detected events, false alarms, and burden. The center alert ledger had 60 rows, which accounted for six centers, two outcomes, and five recall rates. The center alert ledger estimated alert volume at each center based on a common operating point over center denominators. The domain panel had nine rows, which represented five pre-transplant domains and four follow-up domains.

The four-part structure in Table 1 represents the bounds of the manuscript's analysis. The first two panels examine statistical signal and threshold properties; the third panel transforms the threshold value into center-specific review queues; and the final panel examines whether the explanatory inventory is general or focused on a single clinical domain. Reading the panels together ensures that the results do not degenerate into an ordered list of individual performance numbers.

The overall denominator was 4728. Center denominators were 721 in Bern, 790 in Lausanne, 485 in Geneva, 306 in St. Gallen, 1118 in Basel, and 1308 in Zurich. The average recipient age was 51.4 years; 64.5% of the recipient cohort was male; deceased donations comprised 63.1% of the transplants; and 84.1% of the recipients were in their first transplant. Zurich and Basel provided the largest proportions of the cohort, 27.7% and 23.6%, respectively. This difference in denominators becomes significant in the stewardship analysis since common thresholds yield different absolute alert queues.

The denominator breakdown seen in Figure 2 explains why workload planning cannot be based solely on recall and specificity. Since the largest proportion of the transplant-recipient denominator is Basel and Zurich, these centers have the largest absolute alert queues regardless of the shared statistical threshold.

The center summary in Table 2 gives the cohort scale

Table 1. Analysis tables used for multicenter alert stewardship.

Panel	Rows	Main unit	Purpose
Longitudinal model panel	20	Model–outcome–timing	Compares AUROC, AUPRC, longitudinal gain, and prevalence-adjusted enrichment.
Operating-point panel	10	Outcome–recall	Converts threshold operating characteristics into prevalence, alert share, event capture, false reviews, and evaluation burden.
Center alert ledger	60	Center–outcome–recall	Estimates how a common rule distributes alert volume across Bern, Lausanne, Geneva, St. Gallen, Basel, and Zurich.
Domain breadth panel	9	Timing–domain	Measures whether apparent predictor diversity remains broad after domain concentration is considered.

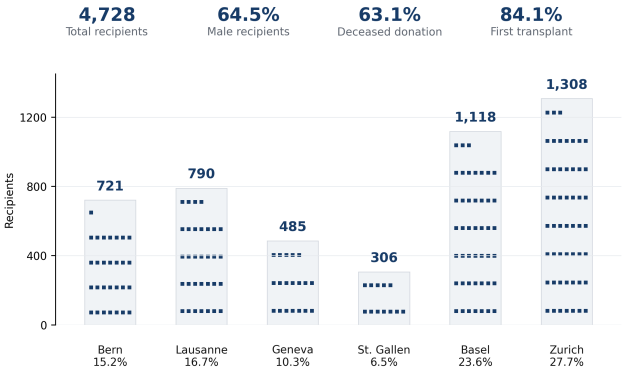


Figure 2. Center denominators and cohort share.

and donor-recipient characteristics needed to interpret later alert-volume differences.

The center table captures the denominator reasoning for the multi-center calculations. Both Zurich and Basel centers are not mere descriptors but define how a seemingly small flagging rate translates into a bigger number of actual chart reviews. For smaller centers, on the contrary, the total number of triggers can be lower but yearly fluctuation in observed precision can be greater due to low numbers of captured events. That is why the ledger includes both network-level and center-level measures.

3 Methods

In this section we describe DCDB and MASA calculations used in the paper. No extra predictive modeling, resampling approach, feature selection procedure, or optimization algorithm is described. Every equation converts one particular aggregate measure into another needed for clinical interpretation: enrichment, workload, center-specific alert count, threshold enlargement, or domain size. This difference matters since the paper assesses the deployability of STCS prognostic tables and not a new modeling approach.

3.1 Decision-calibrated domain balancing

The first layer of analysis used DCDB. Let m be a model, o be an outcome, and t be the timing stage. Follow-up AUPRC is $P^F_{m,o}$, follow-up AUROC is $A^F_{m,o}$, and the respective pre-transplant values are $P^B_{m,o}$ and $A^B_{m,o}$. Follow-up event incidence is π_o. Since AUPRC is highly sensitive to the value of event incidence, DCDB uses prevalence-adjusted ranking enrichment defined as

$$\rho_{m,o} = \frac{P^F_{m,o}}{\pi_o}. \tag{1}$$

This ratio represents the gap between the positive rank ordering and the rate at which positive events are ranked relative to chance. Values closer to 1.0 suggest performance levels close to those seen at the prevalence level, while high values represent increasingly strong association of true positives with high rankings. This ratio is particularly helpful in situations where incidences of the two outcomes may be different since they can lead to different AUPRC values in each case [19, 20].

Longitudinal gain was calculated as

$$\Delta A_{m,o} = A^F_{m,o} - A^B_{m,o}, \tag{2}$$

$$\Delta P_{m,o} = P^F_{m,o} - P^B_{m,o}. \tag{3}$$

These two parameters distinguish progress towards better global discrimination from better positive-class ordering. This is clinically significant since the annual surveillance of transplantation involves identifying the relatively few patients for whom re-review might prevent graft dysfunction.

The recall target evaluation effort at recall target r was defined as

$$E_{o,r} = \frac{1}{Q_{o,r}}, \tag{4}$$

Where $Q_{o,r}$ is the precision. This measure approximates the number of recipients that need to be

Table 2. Center denominators and selected cohort characteristics.

Center	Cohort size	Cohort share (%)	Recipient age, mean	Donor age, mean	Male recipients (%)	Deceased donation (%)
Bern	721	15.2	51.2	54.3	64.2	65.7
Lausanne	790	16.7	52.0	51.9	67.3	58.5
Geneva	485	10.3	52.8	52.8	63.1	57.9
St. Gallen	306	6.5	52.4	55.5	67.3	70.9
Basel	1118	23.6	52.7	52.6	65.7	56.5
Zurich	1308	27.7	49.2	50.6	61.9	70.1
Network	4728	100.0	51.4	52.4	64.5	63.1

reviewed for each detected event. Therefore, a graft loss rule with precision 0.158 needs 6.3 reviews for each captured event while a death rule with precision 0.048 needs 20.8 reviews for each captured event. This index does not substitute the decision curve analysis as it measures the net benefit at different probability threshold [22]. Its function is more specific and operational in nature as it converts the precision in an understandable term for clinicians.

The domain saturation was measured for each timing period. Let $v_{g,t}$ be the number of variables in domain g at time period t. Then the nominal share of domain is

$$D_{g,t} = \frac{v_{g,t}}{\sum_h v_{h,t}}. \tag{5}$$

This measure helps understand whether there is numerical domination of one domain within an inventory of predictors. Domain saturation is important because feature-importance measures may overestimate importance of certain variable families that include many similar variables despite describing a common biological process. SHAP and other interpretability techniques are helpful in interpreting clinical models [26], but clinical governance needs domain analysis too.

3.2 Prevalence recovery from operating points

MASS starts with recovering the event prevalence implied by each operating point. We denote recall as $R_{o,r}$, precision as $Q_{o,r}$, and specificity as $S_{o,r}$. Implied prevalence is

$$\widehat{\pi}_{o,r} = \frac{Q_{o,r}(1 - S_{o,r})}{R_{o,r}(1 - Q_{o,r}) + Q_{o,r}(1 - S_{o,r})}. \tag{6}$$

This is derived from the sensitivity, specificity, and precision formulae. If the recovered prevalence does not change significantly as the recall target changes for the same outcome, then the operating point table has consistency within itself. If there are significant differences in the recovered prevalence, the researcher needs to look at rounding, resampling, or threshold specific sample composition when making the workload inference. In this study, the recovered prevalence is evaluated against the Fan et al. performance table where follow-up incidence was 2.79% for death and 1.49% for graft loss [18].

3.3 Network and center level alert ledger

For each outcome and recall target, the MASA algorithm determines the fraction of recipients that would receive alerts:

$$\widehat{\phi}_{o,r} = R_{o,r}\widehat{\pi}_{o,r} + (1 - S_{o,r})(1 - \widehat{\pi}_{o,r}). \tag{7}$$

For a center c with denominator n_c, the expected number of flagged recipients is

$$\widehat{A}_{c,o,r} = n_c\widehat{\phi}_{o,r}. \tag{8}$$

The expected number of captured events is

$$\widehat{T}_{c,o,r} = n_c R_{o,r}\widehat{\pi}_{o,r}, \tag{9}$$

and expected false reviews are

$$\widehat{F}_{c,o,r} = \widehat{A}_{c,o,r} - \widehat{T}_{c,o,r}. \tag{10}$$

These models make no assumption of uniform risk across centers. These calculations give the planning estimates under the overall operating characteristics for the network. Individual implementation would necessitate center-specific calibration, especially when there are differences among subgroup or centers regarding incidence of events, data quality, immunologic risk, donor mix, or follow-up effort [27].

3.4 Threshold expansion and effective breadth of the domain

The magnitude of an increase in recall may be small but result in a substantial increase in the number of identified patients. The expansion ratio between two recall thresholds r_1 and r_2 defined by MASA is

$$\mathrm{TEF}_{o,r_1 \to r_2} = \frac{\widehat{\phi}_{o,r_2}}{\widehat{\phi}_{o,r_1}}. \tag{11}$$

A 2.0 value indicates that changing one recall procedure to another increases alerts by a factor of two. This variable is an essential stewardship tool due to the fact that healthcare practitioners generally tend to perceive such changes on an absolute basis and not in terms of their effect on coverage.

Effective domain breadth was determined through inverse Herfindahl concentration index:

$$B_t = \left(\sum_{g=1}^{G_t} D_{g,t}^2 \right)^{-1}. \tag{12}$$

The concentration penalty is

$$\kappa_t = 1 - \frac{B_t}{G_t}, \tag{13}$$

where G_t represents the count of nominal domains at time t. A high penalty suggests that the collection of predictors includes a few domains but acts as if it included less because one or two categories have taken over the total number of variables. The measure aids in clear model explanation and is consistent with the clinical prediction system reporting framework which prioritizes interpretability, fairness, and oversight over optimizing performance alone [28].

3.5 Stewardship admissibility rule

The recall target was considered a stewardship-admissible one in the case of satisfaction of three conditions:

$$S_{o,r} \geq 0.85, \tag{14}$$

$$E_{o,r} \leq 12, \tag{15}$$

$$100\widehat{\phi}_{o,r} \leq 15. \tag{16}$$

The first criterion ensures specificity, the second restricts the number of assessments per alerting incident, and the third limits the network alert load. All of these cutoffs are deliberately clear-cut and can be modified by a particular transplant program. For instance, a transplantation center that possesses a safe approach to alerting response, well-developed nurse assessment capabilities, or greater tolerance of false alerts will set a higher limit on alerts. However, centers that conduct invasive investigation due to alerts will have to establish more stringent limits. Clinical guidelines for management of transplant patients call for prolonged monitoring and personalized follow-up, yet predictive alerting requires further adjustments.

4 Results

4.1 Model performance, enrichment, and longitudinal gain

The longitudinal model panel showed the improvement from follow-up data to be quite substantial when comparing against pre-transplant-only modeling. With regards to death, TabPFN provided the greatest follow-up AUROC of 0.802 and follow-up AUPRC of 0.124, while LightGBM performed almost equally with AUROC of 0.797 and AUPRC of 0.122. In the case of graft loss, LightGBM delivered the highest follow-up AUROC of 0.896 and AUPRC of 0.276. TabPFN delivered AUROC of 0.857 and AUPRC of 0.236 for graft loss, while logistic regression delivered AUROC of 0.859 and AUPRC of 0.225. All this is consistent with the findings of Fan et al. that LightGBM would be the most suitable model for further interpretation, because it combines good performance, interpretability, and ease of implementation [18]. Such a result is consistent with clinical experience with machine learning where gradient boosting is known to deliver high quality results on structured clinical data, while tabular foundation models can compete in the case of smaller and mixed tabular datasets [29, 30].

The prevalence-adjusted ranking enrichment provided the insight into interpretation of the rare event problem. With regards to death, the values for TabPFN and LightGBM were similar, at 4.44 and 4.37, respectively. In the case of graft loss, the enrichment value for LightGBM was substantially higher, reaching 18.52, while for TabPFN it was 15.84, while for logistic regression it was 15.10. Thus, not only AUROC, but also ranking enrichment in relation to event incidence was a key point in determining the model suitability for deployment.

The quantities for the DCDB model are presented in Table 3, allowing all three important metrics – global discrimination gain, positive-class ranking gain, and incidence-normalized yield – to be considered simultaneously.

Table 3. DCDB model enrichment and longitudinal gain calculated from the Fan et al. performance table.

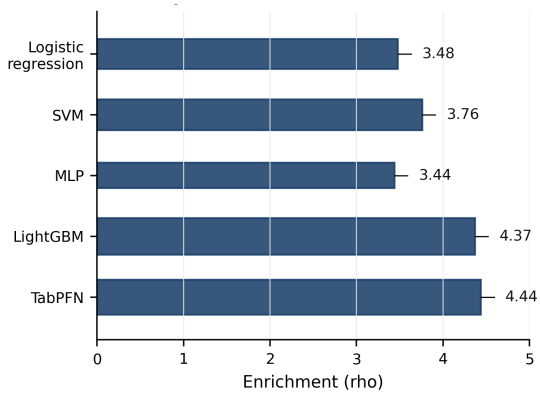
Model	Death			Graft loss		
	ΔAUROC	ΔAUPRC	Enrichment ρ	ΔAUROC	ΔAUPRC	Enrichment ρ
Logistic regression	0.154	0.061	3.48	0.263	0.158	15.10
Support vector machine	0.121	0.063	3.76	0.193	0.154	14.09
Multilayer perceptron	-0.008	0.046	3.44	0.221	0.125	12.35
LightGBM	0.153	0.081	4.37	0.212	0.192	18.52
TabPFN	0.113	0.076	4.44	0.193	0.169	15.84

The table distinguishes two ways in which

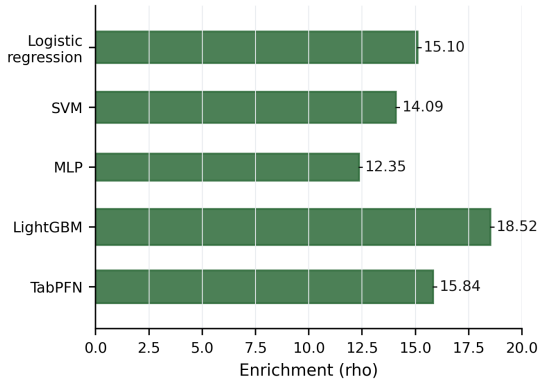
improvement may be misunderstood as the same. The AUROC gain represents improved global ordering, while the AUPRC gain and enrichment represent improved concentration of rare events among the high-risk patients. The graft-loss enrichment is higher in the latter, thus, justifying that the downstream MASA test focuses on graft-loss alerting, before death prediction is considered as a general pathway for follow-ups.

LightGBM produced a greater longitudinal gain for graft loss, increasing the AUPRC for graft loss by 0.192 (more than twice the increase in death AUPRC). Logistic regression produced the greatest AUROC gain for graft loss (0.263), but LightGBM produced the better positive-class ordering. For death prediction, the multilayer perceptron resulted in a negative AUROC gain, even though there was a positive AUPRC gain. The difference between the models' gains justifies the two-stage process: DCDB initially picks the model-outcome pairs with high enough enrichment, and then MASA analyzes whether the thresholds are practically feasible.

By comparison, the yield plane diagram shown in Figure 3, the model-outcome comparison becomes more clear than a usual performance table. While death prediction takes a narrow range of enrichments, graft-loss prediction has a wider range of incidences-adjusted yields, with LightGBM showing the clearest positive-class ordering signal.



(a) Death.



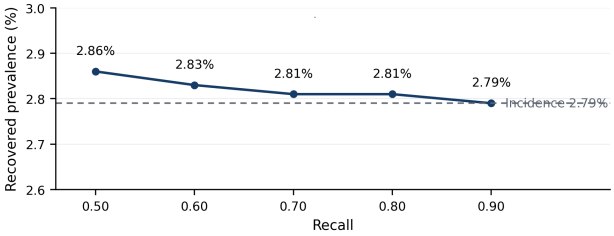
(b) Graft loss.

Figure 3. Incidence-normalized model yield.

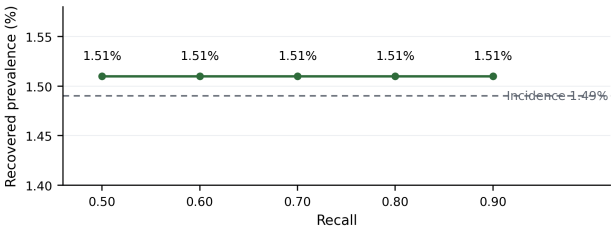
4.2 Recovered prevalence and total network alert volume

The recovered prevalence values are stable in each outcome class. For mortality, the recovered prevalence is stable between 2.79% and 2.86%, with a mean of 2.82%. For graft loss, the recovered prevalence is stable between 1.51% and 1.51%, with a mean of 1.51%. This is consistent with the STCS incidence proportion of 2.79% for mortality and 1.49% for graft loss. The recovered values being stable suggests that the use of the operating point table in terms of aggregate alert volume is justified.

The prevalence gauges in Figure 4 suggest that the recovered prevalence values are near the STCS incidence references regardless of recall target. This means that the use of the operating point table is justified in terms of estimated alert volume without having to use the individual confusion matrices.



(a) Death.



(b) Graft loss.

Figure 4. Recovered prevalence across recall targets.

The full operating-point conversion appears in Table 4, where each recall target is translated into the expected flagged share, network alert count, and captured-event count.

The operating ledger marks the point where the

Table 4. MASA operating-point ledger for network-wide alert volume using the full cohort denominator.

Outcome	Recall	Precision	Specificity	Recovered prevalence (%)	Flagged share (%)	Flagged recipients	Captured events
Death	0.50	0.104	0.873	2.86	13.77	651.0	67.7
Death	0.60	0.083	0.807	2.83	20.45	966.9	80.3
Death	0.70	0.071	0.735	2.81	27.72	1310.8	93.1
Death	0.80	0.058	0.625	2.81	38.69	1829.4	106.1
Death	0.90	0.048	0.487	2.79	52.38	2476.6	118.9
Graft loss	0.50	0.228	0.974	1.51	3.32	156.8	35.8
Graft loss	0.60	0.158	0.951	1.51	5.73	271.0	42.8
Graft loss	0.70	0.097	0.900	1.51	10.91	515.7	50.0
Graft loss	0.80	0.047	0.751	1.51	25.73	1216.7	57.2
Graft loss	0.90	0.039	0.659	1.51	34.95	1652.3	64.4

model threshold turns into a service quantity. The precision and the specificity are still apparent, but the extra columns provide a clinical meaning: a program can estimate how many chart reviews, laboratory evaluations, and clinical records would be needed in order to identify the predicted amount of events. It becomes especially relevant for the rare outcomes when a small change in specificity dominates the alert queue.

The network ledger altered the meaning of the recall. Recall of death equal to 0.70 would have identified more events than recall of 0.50, but at the same time it would have produced almost 660 additional alerts in the denominator of the network. Recall of death equal to 0.90 would have flagged more than half of the recipients in the denominator of the cohort. This recall could not be used as the alert pathway unless the alert response was extremely cheap. Graft loss recall was more acceptable, but it came at a price. Increasing the death recall from 0.60 to 0.70 would have increased the number of identified events from 42.8 to 50.0 and flagged recipients from 271.0 to 515.7. Therefore, seven additional events required 245 additional reviews. This did not make the higher recall rule invalid; it provided a capacity level when the use of this rule became acceptable.

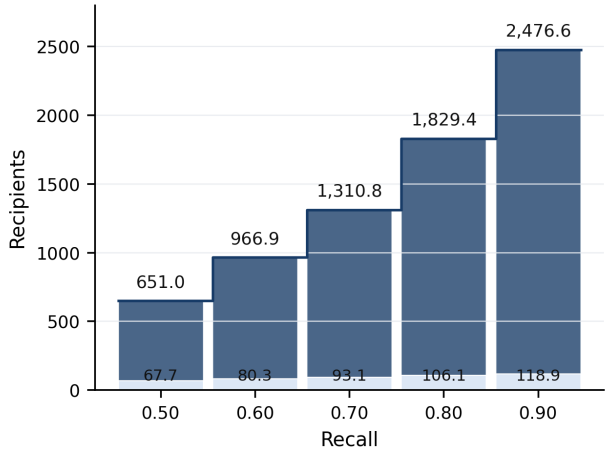
The alert queue staircase presented in Figure 5 illustrates the trade-off directly. Number of the captured events increases gradually, while the number of flagged recipients increases much faster, especially for the high-recall death rules and graft loss recall above 0.70.

The recovered-prevalence calculations also provided a numerical quality check. Death went from a slight movement from 2.86% at recall 0.50 to 2.79% at recall 0.90 – consistent with rounding of the published precision and specificity. Graft loss, on the other hand, showed near-perfect recovered prevalences at all recall targets evaluated. This is important because the MASA approach requires that the recovery be stable and internally coherent with respect to recall, precision and specificity values. Had there been variability in the recovered prevalences at different thresholds, the alert-estimates would have been tagged unstable and the methodology would have needed to be supplemented either with access to the individual confusion matrices or the denominator specific to each threshold. In this case, the operating table is coherent enough for use on deployment scale.

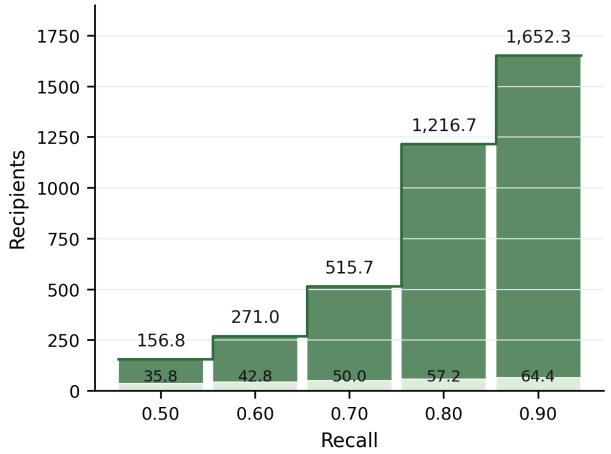
The evaluation burden numbers further explain why graft loss prediction has a better-defined first use case. At recall 0.50, the number of evaluations required per captured event was approximately 4.4, and at recall 0.60 it rose to approximately 6.3 evaluations. These numbers are in a reasonable range for a transplant program given that the first-response can include structured review and confirmatory testing. For death prediction, the number of evaluations required was approximately 9.6 at 0.50 recall and increased monotonically as the recall was increased. While death is a more important endpoint, the relatively low precision suggests that a broad-based death alert program will have to be designed carefully to prevent saturation of routine follow up capacity. The outcome suggests that death prediction should serve as a tool for structured review rather than the initial alert pathway.

False-review counts provide additional illustration of the stewardship challenge. At 0.50 death recall, out of 651 alerted recipients, approximately 583 would be false positives under the recovered aggregate operating characteristics. At 0.60 graft loss recall, out of 271 alerted recipients, approximately 228 would be false positives. This is normal behavior in a rare-event

setting and should not be interpreted as an indication of a poor-performing model. The result, however, demonstrates the need for careful matching between alert precision and clinical response. Review and confirmation through repeat lab tests may be appropriate for a lower-precision alert pathway but a direct course to investigation is not.



(a) Death.



(b) Graft loss.

Figure 5. Network alert queues by recall target.

4.3 Distribution of center alerts

The center alert log revealed that one threshold does not mean an equal work load. Since Zurich and Basel had the largest denominators, their absolute number of alerts was the largest according to every admissible rule. Thus, at a 0.50 death recall, Zurich would have about 180 alerts while Basel would have about 154, as compared to only 42 for St. Gallen and 67 for Geneva. For 0.60 graft-loss recall, Zurich would have 75 alerts while Basel would have 64. With increasing graft-loss recall to 0.70, Zurich's alert prediction would increase to about 143 and Basel's to about 122.

The center alert log in Table 5 utilizes all the candidate rules for all the center denominators.

Table 5. Center-level alert estimates for candidate deployment rules.

Center	Cohort size	Death recall 0.50		Graft-loss recall 0.60		Graft-loss recall 0.70 alerts
		Alerts	Captured events	Alerts	Captured events	Alerts
Bern	721	99.3	10.3	41.3	6.5	78.6
Lausanne	790	108.8	11.3	45.3	7.2	86.2
Geneva	485	66.8	6.9	27.8	4.4	52.9
St. Gallen	306	42.1	4.4	17.5	2.8	33.4
Basel	1118	153.9	16.0	64.1	10.1	121.9
Zurich	1308	180.1	18.7	75.0	11.8	142.7

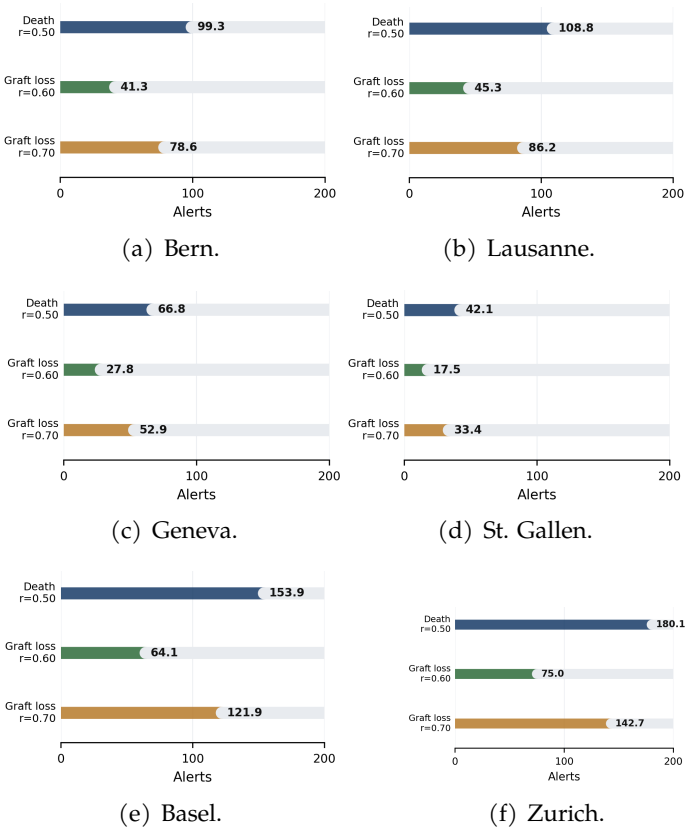
This means that for the site-level deployment, it would be unwise to rely solely on the proportional operating characteristics. Both Zurich and Basel would require a staffing plan to deal with a significantly higher alert queue than St. Gallen, despite the same threshold being used at all centers. However, both St. Gallen and Geneva would get fewer alerts, but with the lower number of captured events, the local estimates of precision would become more volatile. Thus, the governance committee of the network could leverage MASA in order to evaluate the expected number of alerts before the deployment and compare this estimate with the actual number after the deployment in order to determine whether the center has enough difference in the number of events or in the data structure to necessitate a local calibration.

The differences described above imply some practical measures. Network-level thresholds can be set to the same numerical value across the centers, but each center still requires its own operating plan. Larger centers will need a higher number of events to be included in the review queue, additional efforts from the staff and more structured audit processes. Small centers may receive fewer alerts, but with more volatility in local precision monitoring because of the low number of true events. The centrally developed rule, thus, needs to be complemented with local dashboards tracking the alert volume, number of captured events, false positives and missed events.

As the center queue cards in Figure 6 show, the standard rule may be numerically the same across all centers, but the practical review queue is largest in Zurich and Basel due to their large denominators.

4.4 Threshold expansion

Threshold expansion distinguished between statistically plausible recall changes and operationally plausible recall changes. Recall at death increased from 13.77% of the denominator of the entire cohort at recall 0.50 to 52.38% at recall 0.90, an expansion of



(a) Bern. (b) Lausanne.

(c) Geneva. (d) St. Gallen.

(e) Basel. (f) Zurich.

Figure 6. Center-specific alert queues.

3.80 times. The shift from 0.50 to 0.60 recall for death increased the alert share by 1.49 times and resulted in an evaluation effort of about 12.0 evaluations per event detected. For graft loss, the shift from 0.60 to 0.70 recall increased the alert share by 1.90 times, and the shift from 0.70 to 0.80 recall increased the alert share by another 2.36 times. The 0.70 to 0.80 transition is particularly relevant since it results in alert shares above 15%, bringing the specificity down to 75.1%.

The threshold expansion numbers in Table 6 quantify how many additional evaluations will be generated by the specified recall changes.

The implications of these results are that it makes little sense to view recall as a purely linear clinical decision. Recall changes of ten points will result in a much bigger proportionate change in workload with reductions in specificity. This is especially important for transplant clinics, where the consequences of an alert include more than just notification of the incident; there may be a need for nephrologist involvement, transplant-coordinator contact, lab work, immunologic studies, or multidisciplinary conference discussions. Thus, the concept of an expansion factor is useful in turning a statistical threshold into an issue of manpower and policy.

Table 6. Threshold expansion factors for clinically relevant recall changes.

Outcome	Recall change	Flagged-share change	Expansion factor
Death	0.50 to 0.60	13.77% to 20.45%	1.49
Death	0.50 to 0.90	13.77% to 52.38%	3.80
Graft loss	0.60 to 0.70	5.73% to 10.91%	1.90
Graft loss	0.70 to 0.80	10.91% to 25.73%	2.36
Graft loss	0.60 to 0.90	5.73% to 34.95%	6.10

The expansion ruler illustrated in Figure 7 is designed to highlight the operational significance of recall changes. For example, graft-loss recall changes from 0.60 to 0.70 are nearly a doubling of alerts, while recall changes from 0.60 to 0.90 involve sixfold expansion.

4.5 Domain-effective breadth

The STCS predictor table had 51 pre-transplant predictors spanning five nominal domains, while there were 47 follow-up predictors spread over four nominal domains. The pre-transplant list was skewed towards immunologic and HLA predictors, which comprised 30 out of 51 predictors, or 58.8%. The follow-up list was skewed towards rejection and histopathology predictors, which constituted 24 out of 47 predictors, or 51.1%. This skewing is realistic in a clinical context because immunologic compatibility and rejection phenomena are critical in transplantation. However, in terms of explanation governance, nominal domain number overestimates predictor inventory diversity.

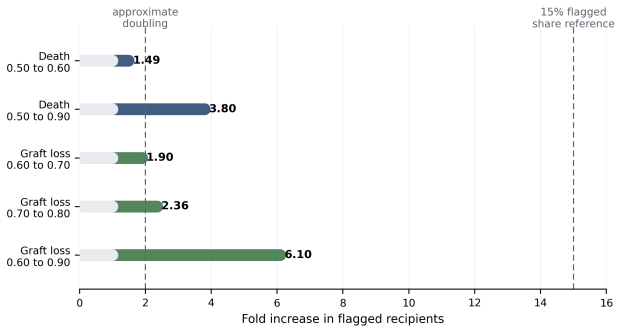


Figure 7. Threshold expansion of flagged recipients.

Table 7 displays the diversity measures, comparing nominal domain number with effective diversity based on concentration.

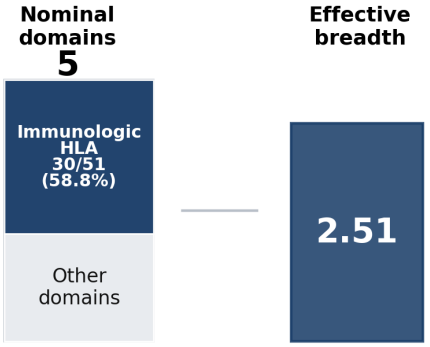
The pre-transplant prediction predictor inventory was quite broad at 2.51 effective domains despite its five nominal domains. The concentration penalty was equal to 0.50, implying that the effective explanatory

Table 7. Predictor-domain saturation and effective breadth.

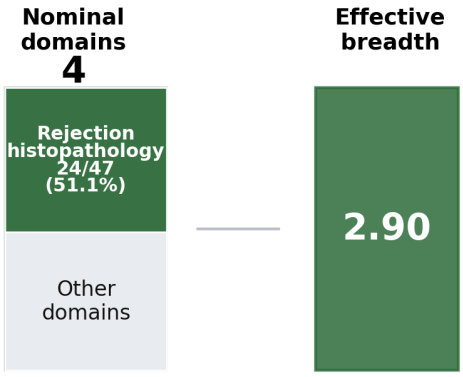
Timing	Dominant domain	Variables	Largest domain share (%)	Effective breadth	Concentration penalty
Pre-transplant	Immunologic and HLA	51	58.8	2.51	0.50
Follow-up	Rejection and histopathology	47	51.1	2.90	0.28

breadth is about half of the nominal breadth. The follow-up predictor inventory was broad at 2.90 effective domains even though the inventory had four nominal domains resulting in a concentration penalty of only 0.28. This suggests that the follow-up prediction, which is still dominated by rejection and histopathology predictors, uses a wider clinical base of prediction as compared to pre-transplant prediction. This finding favors grouped interpretative reports that use renal function, immunological variables, rejection and biopsy results, infections, comorbidities, and CKD stages as the clinical domains.

The domain-compression view illustrated in Figure 8 clearly visualizes this concentration phenomenon. The pre-transplant predictor inventory is rather wide when considered based on its nominal breadth, but immune system and HLA predictors compress the effective breadth to 2.51 domains; the follow-up inventory stays relatively distributed when adjusted for rejection and histopathology dominance.



(a) Pre-transplant inventory.



(b) Follow-up inventory.

Figure 8. Effective predictor-domain breadth.

4.6 Admissible stewardship classification

The last stewardship classification used both DCDB and MASA. The recall for graft loss of 0.60 met all three stewardship criteria: the specificity was 95.1%, the evaluation burden was 6.3 evaluations per captured event, and the flagged share was 5.73%. Recall for graft loss of 0.70 also met the specificity, burden, and flagged share thresholds, but it doubled the network alert generation in comparison to recall of 0.60. Recall for death of 0.50 also met the criteria, but it produced higher network alert generation compared to the recall of graft loss of 0.60 and also had lower model enrichment. Recall for death of 0.60 failed the flagged share threshold and approached the burden threshold. Recall for graft loss of 0.80 failed due to low specificity of 75.1%, increased burden of 21.3, and raised flagged share of 25.73%.

The last admissible release-track chart shown in Figure 9 compares all the candidate rules according to the same three tests. It should be noted that the preferred starting rule stands out in terms of visualization as it possesses high specificity, low number of evaluations per captured event, low flagged share, and comparatively small network alert generation.

The candidate-rule comparison presented in Table 8 is a combination of the following metrics: specificity, evaluation load, flag rate, alert rate, and interpretation.

Threshold deployment is thus more informative than performance alone since a specific combination of parameters allows a more precise decision recommendation. Graft-loss prediction can be viewed as the preferable initial task since it is enriched better and involves less alerts. The 0.60 rule for graft loss appears to be more conservative choice. While the

Table 8. Final stewardship interpretation after combining DCDB and MASA.

Rule	Specificity (%)	Evaluations per event	Flagged share (%)	Network alerts	Interpretation
Death recall 0.50	87.3	9.6	13.77	651.0	Usable but operationally heavy
Death recall 0.60	80.7	12.0	20.45	966.9	Not preferred because alert share is high
Graft-loss recall 0.60	95.1	6.3	5.73	271.0	Preferred starting rule
Graft-loss recall 0.70	90.0	10.3	10.91	515.7	Capacity-dependent escalation
Graft-loss recall 0.80	75.1	21.3	25.73	1216.7	Not admissible

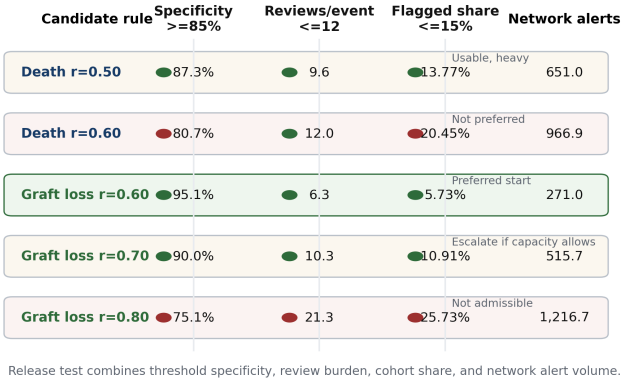


Figure 9. Release track admissibility classification.

0.70 rule is not excluded, it is rather a capacity-based upgrade since it detects around seven extra events while generating around 245 extra alerts in the whole cohort denominator. Death prediction should also be regarded as clinically meaningful, yet should be applied cautiously with some indication narrowing, clinical screening, or long-term event prediction.

5 Discussion

The research question was whether aggregated threshold tables in a kidney-transplantation prognostic study could be transformed into the center-level review ledger which would allow setting up a feasible starting rule for next-year death and graft-loss surveillance. The answer is affirmative only if model performance is considered in light of alert volume, center scale, threshold expansion, and explanation governance. There is enough data in STCS prognostic tables to switch from model comparison to deployment planning without introducing a new model. This is the key benefit compared to the traditional approach to interpretation of predictive models' performance.

The first important finding is that graft-loss prediction is the more governable task. LightGBM yielded the strongest enrichment in terms of AUPRC-to-incidence ratio – 18.52, while death models showed relatively low values of this metric. This does not mean that death prediction is clinically useless. Death is a vital endpoint and can be helpful for risk discussions, cardiovascular review, frailty assessment, infections surveillance, and multidisciplinary planning. However, a wide one-year death alert involves much review traffic per each detected event. In the course of developing a structured alerting pathway in a transplantation network, graft loss is thus the better endpoint because of better rare-event concentration and association with transplantation-specific surveillance activities.

The second important finding is that recall is not service-neutral. The change from 0.60 to 0.70 recall rate for graft loss results in increasing captured events from about 43 to 50 but in expanding flagged recipients from about 271 to 516. The additional event capture is clinically valuable, but the review queue almost doubles. This trade-off explains why the 0.70 threshold cannot be simply called better or worse. It can be regarded as a capacity-based escalation. In case of a brief chart review or automatic lab reminder, the 0.70 rule can be used. In case of a nephrologist review, immunologic testing, coordinator consultation, and multidisciplinary discussion, the 0.60 rule will be the safer choice. The MASA ledger highlights this point clearly.

The third finding is that the center denominators matter for implementation burden even if statistical performance is the same. Zurich and Basel have the largest absolute number of alerts due to their largest cohort denominators. Other centers have the smaller number of expected alerts, but may have the higher instability of observed precision because of sparsity of captured events. The finding has a direct impact on governance: center-level monitoring must include local reports on the alert volume and observed yield, while precision and number of missed events assessment will require network-level pooling over time. The finding also supports staged implementation in which the preferred network rule is adopted first and center-level deviation is analyzed after collecting enough alerts.

The fourth finding is related to the explanation

governance. The pre-transplant inventory includes five nominal domains, but the real number of domains is 2.51 since the immunologic and HLA variables are dominating in the total number. The follow-up inventory includes four nominal domains and demonstrates higher number of domains equal to 2.90 despite a dominating rejection and histopathology part. This point is important because the clinical user can expect that a greater number of named domains indicates broader evidence. The domain breadth calculation shows that the nominal diversity can exaggerate the effective diversity. Reporting by groups on the renal function, rejection, histopathology, immunologic activity, infection, comorbidities, and treatments will thus allow providing a more clinically interpretable explanation than a long unstructured feature list.

The figures and tables serve complementary purpose in this interpretation. The analytical anatomy chart provides the structure of the evidence pathway from STCS tables to admissible alert rules. The denominator and center queue charts show that the implementation is not homogeneous across the centers. The model yield and prevalence recovery charts prove that the threshold calculation is numerically correct. The network queue and expansion charts explain why a higher recall target can be turned into a workload problem earlier than into a statistical one. The domain breadth chart shows why the explanation must be understood in terms of clinical domains, and the release track chart finds the rule which fulfills the requirements to specificity, review burden, and alert share simultaneously. These visual and tabular elements keep the manuscript focused on the service aspect of a threshold.

The findings are consistent with the general guidance for clinical prediction and clinical machine learning. Transparent reporting and structured bias appraisal remain a necessity if a prediction model is connected to clinical service [24, 25]. The calibration assessment is also required before using the risk estimates in practice [21]. Responsible clinical machine learning also requires post-deployment monitoring and explicit ethical governance [27, 28]. This paper contributes a narrow practical step of translation for the cases when a secondary analyst cannot get access to the patient-level data. If recall, precision, specificity, event incidence, and center denominators are provided, a reader can make a rough estimation of the threshold governability before a prospective deployment.

Several boundaries must be distinguished. The analysis is performed with aggregate tables and does not substitute the patient-level calibration, external validation, fairness assessment, or prospective clinical evaluation. The center-level precision may differ from the network-level estimate since the local case mix, missingness, practice patterns, follow-up timing, and event incidence are different across the centers. The use of the whole cohort denominator is the planning convention; the true number of annual alerts will depend on the number of event-free recipients still under surveillance at each prediction time. The thresholds admissibility cutoffs for specificity, evaluations per event, and flagged share are transparent policies, not universal constants. Finally, the effective number of domains is calculated in terms of variable counts, not weighted model contribution. A future implementation study can be based on the same ledger with the observed alert volume, calibration in the large, grouped SHAP values, decision curve net benefit, and prospective safety monitoring.

6 Clinical and Methodological Implications

For clinical use, the analysis favors a sequential protocol. A transplant network should initially select model-outcome pairs with significant rare-event enrichment and translate candidate threshold values into alert share, flagged events, false reviews, and workload at individual centers. Next, the network should evaluate observed precision, event misses, workload, and center variance against pre-deployment benchmark values. The order of operations ensures that the model is not deployed based on its position within the model performance table alone.

For kidney transplant follow-up, the most conservative starting point is graft loss recall 0.60. This threshold preserves the specificity value at 95.1%, highlights 5.73% of the total number of transplantations, requires 6.3 evaluations per event detected, and creates the smallest possible admissible alert lists at centers. Graft loss recall 0.70 still represents a clinically viable option but needs explicit capacity approval due to the 1.9-fold increase of the alert queue size compared to 0.60. Prediction of death may find its place in specialized processes, but not in routine follow-up unless the network can handle a much heavier review workload.

As for methodological presentation of results, the paper suggests a simple recommendation: models developed for the purposes of clinical process must provide threshold-specific recall, precision, specificity, predicted alert share, prevalence of outcome, and center-level denominators whenever possible. These metrics are not supposed to replace discrimination and calibration, but they give an opportunity for editors, clinicians, and process managers to judge whether the model could be used as a part of a service process.

7 Conclusion

The paper posed a hypothesis of whether summary tables for aggregate risk models for kidney transplant follow-up could be converted into a center-level deployment ledger for identifying the defensible alert threshold. The hypothesis is confirmed by the analysis of the STCS summary tables. DCDB found the graft loss to be a better choice for further development because LightGBM provided a high prevalence-adjusted enrichment and significant longitudinal gain. Finally, MASA identified graft loss recall 0.60 as the most conservative starting point for deployment because it provides high specificity, low alert share, moderate evaluation burden, reasonable center queue sizes, and interpretable scope of predictions.

The message is not that higher recall is unappealing, but it increases the burden on the reviewers and makes prediction less governed due to quick rise of this burden. Therefore, graft loss recall 0.70 is still an admissible threshold for prediction but it is capacity-based because it nearly doubles the number of alerts. Prediction of death is a clinically important task but the usage of the resulting alerts in routine processes is not very governable due to the same reasons.

References

[1] Loupy, A., & Lefaucheur, C. (2018). Antibody-mediated rejection of solid-organ allografts. *New England Journal of Medicine, 379*(12), 1150-1160.

[2] Nankivell, B. J., & Kuypers, D. R. (2011). Diagnosis and prevention of chronic kidney allograft loss. *The Lancet, 378*(9800), 1428-1437.

[3] Van Loon, E., Bernards, J., Van Craenenbroeck, A. H., & Naesens, M. (2020). The causes of kidney allograft failure: more than alloimmunity. *A viewpoint article. Transplantation, 104*(2), e46-e56.

[4] Coemans, M., Süsal, C., Döhler, B., Anglicheau, D., Giral, M., Bestard, O., ... & Naesens, M. (2018). Analyses of the short-and long-term graft survival after kidney transplantation in Europe between 1986 and 2015. *Kidney international, 94*(5), 964-973.

[5] Loupy, A., Aubert, O., Orandi, B. J., Naesens, M., Bouatou, Y., Raynaud, M., ... & Lefaucheur, C. (2019). Prediction system for risk of allograft loss in patients receiving kidney transplants: *international derivation and validation study. Bmj, 366*.

[6] Yoo, K. D., Noh, J., Lee, H., Kim, D. K., Lim, C. S., Kim, Y. H., ... & Kim, Y. S. (2017). A machine learning approach using survival statistics to predict graft survival in kidney transplant recipients: a multicenter cohort study. *Scientific reports, 7*(1), 8904.

[7] Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems, 106*, 97-109.

[8] Naqvi, S. A. A., Tennankore, K., Vinson, A., Roy, P. C., & Abidi, S. S. R. (2021). Predicting kidney graft survival using machine learning methods: prediction model development and feature significance analysis study. *Journal of Medical Internet Research, 23*(8), e26843.

[9] Van Loon, E., Zhang, W., Coemans, M., De Vos, M., Emonds, M. P., Scheffner, I., ... & Naesens, M. (2021). Forecasting of patient-specific kidney transplant function with a sequence-to-sequence deep learning model. *JAMA network open, 4*(12), e2141617.

[10] Schwab, S., Sidler, D., Haidar, F., Kuhn, C., Schaub, S., Koller, M., ... & Swisstransplant Kidney Working Group (STAN) Amico Patrizia Folie Patrick Gannagé Monique Matter Maurice Nilsson Jakob Peloso Andrea de Rougemont Olivier Schnyder Aurelia Spartà Giuseppina Storni Federico Villard Jean Wirth-müller Urs Wolff Thomas. (2023). Clinical prediction model for prognosis in kidney transplant recipients (KIDMO): study protocol. *Diagnostic and prognostic research, 7*(1), 6.

[11] Salaün, A., Knight, S., Wingfield, L., & Zhu, T. (2024). Predicting graft and patient outcomes following kidney transplantation using interpretable machine learning models. *Scientific Reports, 14*(1), 17356.

[12] Paquette, F. X., Ghassemi, A., Bukhtiyarova, O., Cisse, M., Gagnon, N., Della Vecchia, A., ... & Loudiyi, Y. (2022). Machine learning support for decision-making in kidney transplantation: step-by-step development of a technological solution. *JMIR medical informatics, 10*(6), e34554.

[13] Achilonu, O., Obaido, G., Ogbuokiri, B., Aruleba, K., Musenge, E., & Fabian, J. (2024). A machine learning approach towards assessing consistency and reproducibility: an application to graft survival

across three kidney transplantation eras. *Frontiers in Digital Health, 6*, 1427845.

[14] Fabreti-Oliveira, R. A., Nascimento, E., de Melo Santos, L. H., de Oliveira Santos, M. R., & Veloso, A. A. (2024). Predicting kidney allograft survival with explainable machine learning. *Transplant Immunology, 85*, 102057.

[15] Stampf, S., Mueller, N. J., van Delden, C., Pascual, M., Manuel, O., Banz, V., ... & members of the Swiss Transplant Cohort Study. (2021). Cohort profile: the Swiss Transplant Cohort Study (STCS): a nationwide longitudinal cohort study of all solid organ recipients in Switzerland. *BMJ open, 11*(12), e051176.

[16] Frischknecht, L., Deng, Y., Wehmeier, C., de Rougemont, O., Villard, J., Ferrari-Lacraz, S., ... & Swiss Transplant Cohort Study. (2022). The impact of pre-transplant donor specific antibodies on the outcome of kidney transplantation–Data from the Swiss transplant cohort study. *Frontiers in immunology, 13*, 1005790.

[17] Heldal, T. F., Åsberg, A., Ueland, T., Reisæter, A. V., Pischke, S. E., Mollnes, T. E., ... & Jenssen, T. G. (2023). Systemic inflammation early after kidney transplantation is associated with long-term graft loss: a cohort study. *Frontiers in immunology, 14*, 1253991.

[18] Fan, B., Schürch, M., Tian, Y., Mallone, A., Frischknecht, L., Koller, M., ... & Krauthammer, M. (2025). Enhancing post-kidney transplant prognostication: an interpretable machine learning approach for longitudinal outcome prediction. *npj Digital Medicine, 8*(1), 684.

[19] Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).

[20] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*(3), e0118432.

[21] Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., & Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. (2019). Calibration: the Achilles heel of predictive analytics. *BMC medicine, 17*(1), 230.

[22] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565-574.

[23] Riley, R. D., Ensor, J., Snell, K. I., Harrell, F. E., Martin, G. P., Reitsma, J. B., ... & Van Smeden, M. (2020). Calculating the sample size required for developing a clinical prediction model. *Bmj, 368*.

[24] Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Correction: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement (vol 162, pg 55, 2015). *Annals of Internal Medicine, 162*(8), 600.

[25] Moons, K. G., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., ... & Mallett, S. (2019). PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Annals of internal medicine, 170*(1), W1-W33.

[26] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems, 30*.

[27] Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine, 25*(9), 1337-1340.

[28] Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science, 4*(1), 123-144.

[29] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems, 30*.

[30] Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., ... & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature, 637*(8045), 319-326.